To appear in the *Journal of Atmospheric Sciences*, 2000.

Rotation of EOFs by the Independent Component Analysis: Towards a Solution of the Mixing Problem in the Decomposition of Geophysical Time Series

Filipe Aires
Department of Applied Physics, Columbia University, NASA Goddard Institute for Space Studies, New-York, USA

William B. Rossow
NASA Goddard Institute for Space Studies, New-York, USA

Alain Chédin
CNRS Laboratoire de Météorologie Dynamique, Palaiseau, France

Abstract

The Independent Component Analysis is a recently developed technique for component extraction. This new method requires the statistical independence of the extracted components, a stronger constraint that uses higher-order statistics, instead of the classical decorrelation, a weaker constraint that uses only second-order statistics. This technique has been used recently for the analysis of geophysical time series with the goal of investigating the causes of variability in observed data (i.e. exploratory approach). We demonstrate with a data simulation experiment that, if initialized with a Principal Component Analysis, the Independent Component Analysis performs a rotation of the classical PCA (or EOF) solution. This rotation uses no localization criterion like other Rotation Techniques (RT), only the global generalization of decorrelation by statistical independence is used. This rotation of the PCA solution seems to be able to solve the tendency of PCA to mix several physical phenomena, even when the signal is just their linear sum.

1. Introduction

This work concerns methods for the investigation of the physical causes of variability of a dynamical system, e.g., the climate, from observations of its behavior. The observed time series of the system's state can be produced by a mixture of different components representing different physical processes. In the most general case, the time series, $\boldsymbol{x}(j)$, with temporal dimension, N, at a particular spatial coordinate, $j \in \{1,\ldots,M\}$, where M is the spatial dimension, is the result of the mixture of these components, $\boldsymbol{\sigma}(j)$, by an operator $\mathcal{G}$:

$$\boldsymbol{x}(j) = \mathcal{G}(\boldsymbol{\sigma}(j)). \tag{1}$$

In this paper we use lower (upper) case bold letters to indicate vectors (matrices) and we will refer to a particular spatial location as a pixel. We consider decomposition in time (i.e., the observations, $\boldsymbol{x}(j)$, are time series at each pixel, j), but the following discussion would be the same for a decomposition in space.

The goal of the analysis method is to infer the unknown contributing components, $\boldsymbol{\sigma}(j)$, from the observed data, $\boldsymbol{x}(j)$. Often in performing such an analysis, one does not know a priori much about what $\mathcal{G}$ is like, even whether it is linear or not. We introduce

$$\boldsymbol{h} = \mathcal{J}(\boldsymbol{x}) \simeq \boldsymbol{\sigma}, \tag{2}$$

where $\mathcal{J}$ is an estimate of the unknown inverse mapping, $\mathcal{G}^{-1}$, and $\boldsymbol{h}$ is an estimate of the unknown vector $\boldsymbol{\sigma}$. Analysis methods that estimate $\mathcal{J}$ and $\boldsymbol{h}$ are called component extraction techniques.

If $\mathcal{G}$ is nonlinear and cannot be usefully linearized, then we are faced with a component extraction problem that is highly complex for many reasons. First, the definition of a well-adapted nonlinear component extraction model, $\mathcal{J}$, is difficult without some a priori information about the nonlinear mixture model, $\mathcal{G}$. Using generic statistical models for the nonlinear regressions of $\mathcal{J}$ often introduces too many degrees of freedom, which ruins the inference process. Second, the determination of the extracted components is much more complex since the basis functions, $\boldsymbol{g}_i(\boldsymbol{\sigma}(j))$, vary with location, j. Third, the uniqueness of the result and its interpretation in physical-process terms is difficult to demonstrate. Some nonlinear techniques have been developed (*Monahan* 2000) but their main application is to represent complex data in a more compact way (i.e., compression). Here, we are not concerned with this use of such techniques, but with their use to extract "meaningful" (i.e. causal) components to understand how a dynamical system works. Currently, all the "classical" (and most frequently used) extraction methods are linear; the use of nonlinear models in (4) for component extraction is only just beginning. The nonlinear case is beyond the scope of this paper.

One approach to simplify the analysis that may when $\mathcal{G}$ is complex is to linearize (1) using $\boldsymbol{G}(\boldsymbol{\sigma}^0(j))$, the Jacobian matrix of the nonlinear operator $\mathcal{G}$ near a particular state, $\boldsymbol{\sigma}^0(j)$:

$$\Delta\boldsymbol{x}(j) = \boldsymbol{G}(\boldsymbol{\sigma}^0(j)) \cdot \Delta\boldsymbol{\sigma}(j) \tag{3}$$

$$= \boldsymbol{g}_1(\boldsymbol{\sigma}^0(j))\cdot\Delta\sigma_1(j)+\boldsymbol{g}_2(\boldsymbol{\sigma}^0(j))\cdot\Delta\sigma_2(j)+\cdots+\boldsymbol{g}_Q(\boldsymbol{\sigma}^0(j))\cdot\Delta\sigma_Q(j) \tag{4}$$

where the temporal basis functions, $\boldsymbol{g}_1(\boldsymbol{\sigma}^0(j)),\ldots,\boldsymbol{g}_Q(\boldsymbol{\sigma}^0(j))$, which are the columns of the matrix $\boldsymbol{G}(\boldsymbol{\sigma}^0(j))$, are unknown time series describing a fixed dynamical behavior at each pixel, j. Each $\boldsymbol{g}_i(\boldsymbol{\sigma}^0(j))$ is the temporal response to a perturbation $\Delta\sigma_i(j)$ of i^{th} component at pixel, j, when the state is given by $\boldsymbol{\sigma}^0(j)$. For example, in the case where the physical component, i, is an oscillating wave propagating in space, the time series $\boldsymbol{g}_i(\boldsymbol{\sigma}^0(j))$ have the same shape as the source, but with a time delay dependent on the distance between their location at pixel, j, and the source of the wave.

When $\mathcal{G}$ is known a priori to be linear or has been linearized to $\boldsymbol{G}$, the equivalent to (1)-(4) when the time series, $\boldsymbol{x}(j)$, at particular spatial coordinate, $j \in \{1,\ldots,M\}$, is decomposed in time, is:

$$\boldsymbol{x}(j) = \boldsymbol{G}\cdot\boldsymbol{\sigma}(j) = \boldsymbol{g}_1\sigma_1(j) + \boldsymbol{g}_2\sigma_2(j) + \cdots + \boldsymbol{g}_Q\sigma_Q(j), \tag{5}$$

where the temporal basis functions, $\boldsymbol{g}_1,\ldots,\boldsymbol{g}_Q$, which are the columns of matrix $\boldsymbol{G}$, are unknown time series describing a fixed dynamical behavior. In contrast with the nonlinear case, the basis functions, $\boldsymbol{g}_i$, are independent of the geographical location, j. Each $\boldsymbol{g}_i$ could be a signal with a different physical cause operative in a particular geographical region represented by a different component map $\{\sigma_i(j)\ ;\ j = 1,\ldots M\}$. The goal of the analysis is to infer the unknown contributing components, $\boldsymbol{\sigma}$(j), from the observed data, $\boldsymbol{x}(j)$. In the linear case, we write (2) as

$$\boldsymbol{h} = \boldsymbol{J}\cdot\boldsymbol{x} \simeq \boldsymbol{\sigma}, \tag{6}$$

where $\boldsymbol{J}$ is an estimate of the unknown matrix $\boldsymbol{G}^{-1}$ (the superscript $^{-1}$ represents the pseudo-inverse if $\boldsymbol{G}$ is not square) and $\boldsymbol{h}$ is an estimate of the unknown vector $\boldsymbol{\sigma}$. The ability of statistical analysis techniques

to retrieve good estimates, $\boldsymbol{h}$, of the true components, $\boldsymbol{\sigma}$, is highly dependent on the quality of the statistical dataset used (i.e., sufficiently large number of independent examples is needed to sample all the variations involved) and on the technical assumptions that are made about $\boldsymbol{J}$ and $\boldsymbol{h}$.

There are many techniques that have been developed to estimate $\boldsymbol{J}$ and $\boldsymbol{h}$. The one most frequently used by the climate research community today is the Principal Components Analysis (PCA) or Empirical Orthogonal Function (EOF) method (*Lorentz* 1951; *von Storch and Frankignoul* 1998). Sometimes, modifications of this method are used that either apply some other criterion besides maximizing the variance explained by each component or relax the requirement for orthogonal basis functions. These methods use this additional information to rotate, orthogonally or obliquely, the solution of a previous PCA; we refer to this large set of methods as Rotational Techniques (RT) (see *Horel* 1981; *Richman* 1986). We formulate a general linear case and the classical analysis techniques in Section 2. We will show some of the known difficulties of the PCA solution (*Karl et al.* 1982; *Richman* 1986) that have led to the development of the rotational techniques.

The Independent Component Analysis (ICA) is introduced in Section 3. This method is based on information theory and has been recently developed in the context of signal processing studies and of the development of neural coding models (*Jutten and Heraut* 1991; *Atick* 1992; *Bell and Sejnowski* 1995). This technique has now been studied for some time by the statistical analysis research community and many recent applications of the ICA paradigm can be found in the ICA '99 proceedings (*Cardoso et al.* 1999) or (*Hyvärinen and Oja* 2000), but this method has not been used for analysis of climatological observations (see, *Aires et al.* 2000) The two major distinctions between the ICA approach and the classical techniques are:

- The method extracts statistically independent components, even if these components have non-Gaussian probability distribution functions, by making use of higher-order statistics, whereas the PCA or RT approaches use only second-order statistics.
- A linear mixture model is not assumed; any extraction model could be used with the ICA paradigm (*Burel* 1992), which allows the introduction of pertinent a priori information about the mixture model, if it is available.

We will show that the "linear" (this term will be explained in the following) and PCA-initialized form of ICA, described here, performs a rotation of the PCA solution, eliminating the mixing problem: PCA has the tendency to mix several components, even when the signal is just their linear sum. We argue that, because of certain general features of the ICA approach, it is a particularly promising technique for rotations of the PCA solution. Furthermore, a previous study that applies a similar ICA to the analysis of variations of tropical sea surface temperature time series (*Aires et al.* 2000) illustrates its potential to separate a geophysical time series into more meaningful components.

To illustrate most clearly how ICA avoids the component mixing problem, we construct a synthetic dataset, where the true answer to the decomposition problem is known, and apply PCA-initialized Independent Component Analysis to extract components (Section 4). We have deliberately devised a dataset with certain characteristics to challenge whether ICA can separate distinct modes of variability. In particular, the synthetic dataset is formed by a linear sum of components, some of which are well-separated in space and time, some of which overlap and some of which represent teleconnections. This dataset is then created so that it has the structure that linear component extraction techniques search for. We show that, even in the linear case, the PCA technique mixes the components, but that the ICA method performs a rotation to correctly separate the components.

The goals of this paper are then to illustrate some of the problems of classical analyses, even in the simple linear case, and to introduce a new component extraction technique that overcomes these problems, at least in its linear form (Section 5). We compare linear-ICA to PCA to measure the effect of the rotation transformation by the ICA algorithm. We apply these methods to a synthetic dataset, rather than to an actual geophysical dataset, so that the true answer is known independently and so that we can illustrate several specific features of the component mixing problem. We do not use a priori information for the component extraction experiments since we are interested in exploratory techniques, not confirmatory ones, i.e. we wabt a technique to find the correct, but unknown components, not confirm results from other analysis.

2. The Linear Case and Classical Component Extraction Techniques

A common approach for statistical component extraction is to require the decorrelation of the extracted components, in which case the covariance matrix of extracted components $< \boldsymbol{h}^t \cdot \boldsymbol{h} >$ is constrained to be diagonal; but this decorrelation constraint has an infinity of solutions because

$$\boldsymbol{J} = \Theta \cdot \boldsymbol{J}_0, \tag{7}$$

where Θ is any undetermined $Q \times Q$ matrix so that $\Theta^t \cdot \Theta = \boldsymbol{I}_{Q \times Q}$. $\boldsymbol{J}_0 = \Sigma^{-1/2} \cdot \boldsymbol{E}^t$ is a $Q \times N$ matrix with Σ the truncated diagonal matrix of the higher (in decreasing order) eigenvalues of $< \boldsymbol{x}^t \cdot \boldsymbol{x} >$ and $\boldsymbol{E}$ the $N \times Q$ matrix with the associated normalized eigenvectors in the columns.

One particular decorrelation solution is the well-known Principal Component Analysis (PCA) or, in the geophysical community, Empirical Orthogonal Function[1] (EOF), first used in atmospheric sciences by *Lorenz* (1951). In this technique, an additional constraint is added to resolve the indeterminacy of the decorrelation solutions: the successive extracted components have to explain the maximum remaining variance. This solution is given by taking $\Theta = \boldsymbol{I}_{Q \times Q}$ in Equation (7). Depending on which space the PCA is applied to (space, time, frequency, multivariate data, etc), the PCA has also been called Singular Spectrum Analysis (*Broomhead and King* 1986; *Vautard et al.* 1992), Multi-channel Singular Spectrum Analysis (*Vautard et al.* 1997), Extended Empirical Orthogonal Functions (*Korres et al.* 2000), Multivariate Empirical Orthogonal Function (*Xue et al.* 2000), etc.

Three well-known problems arise when using the PCA technique. (1) Even if the mixing of the components is linear as in Eq. (5), the maximum-explained-variance assumption can lead to a different mixing in the extracted components (*Kim and Wu* 1999) as we will be shown here (see Figure 1a for an schematic illustration of this problem in a 2-dimensional case). (2) This mixing problem is also particularly serious when the PCA is applied to data that have more than one component with about the same variance. In this case, the problem is not solvable since any orthogonal rotation of the principal components (i.e., in the space of the "degenerate" eigenvectors) will be a PCA solution (Figure 1b). (3) Since PCA imposes orthogonality on the extracted basis functions, mixing problems also arise when the actual physical basis functions are not orthogonal (Figure 1c). Another problem for the application of the PCA to geophysical data arises from an irregularly spaced grid of pixels that can lead to distorted basis functions (*Karl et al.* 1982).

The PCA assumptions (linearity, orthogonality, maximum variance explained by each successive component) used to resolve the solution indeterminacy are not known, a priori, to be valid for a particular dataset. If these assumptions are not valid, variations that are not physically connected could be artificially mixed together into one extracted component (i.e., the mixing problem). This is the reason why PCA is often used in restricted geographical domains instead of global domains or applied to pre-filtered data to try to isolate a single dominant mode of variation, which PCA can correctly identify. Thus, although PCA is useful for compressing information by **describing** the most variance with the fewest terms in an expansion (as a dimension-reduction/compression technique), it can lead to misinterpretation of physical relationships when used as a component extraction technique.

Rotational Techniques (RT) were introduced (*Horel* 1981; *Richman* 1981), in part, to obtain a more physically interpretable solution and to avoid some of the problems of PCA. In these approaches, an additional constraint of localization, based on the so-called "simple structure" principle, is used to solve the indeterminacy of the decorrelation solutions. The rotation can be orthogonal (the rotation matrix is an orthogonal matrix) or oblique (this constraint is relaxed). There exist many proposed localization criteria: quartimax, varimax, transvarimax, quartimin, oblimax, *etc* (see the review paper of *Richman* (1986) on this subject). Two general distinct classes of RT solution can be distinguished: confirmatory RT where a priori information about the components is available and we want to verify the hypothesis, and exploratory RT where almost no a priori information on the problem is available. We are interested, in this study, in the exploratory case where no a priori information is available. Since there is no general principle for choosing a particular localization criterion from this large set of proposed solutions is available, use if a particular RT method in exploratory mode may be equivalent to introducing a priori information about the localization that may not be well suited to the particular problem.

We describe here the most commonly used criterion

[1] EOF is a specific form of the general PCA were the extracted basis functions are normalized.

for orthogonal rotation:

$$V_\gamma(\boldsymbol{G}) = \sum_{j=1}^{Q}\sum_{i=1}^{N}(G_{ij})^4 - \frac{\gamma}{N}\sum_{j=1}^{Q}\left[\sum_{i=1}^{N}(G_{ij})^2\right]^2, \quad (8)$$

where the constant γ gives a family of rotations, with $\gamma = 1.0$ giving varimax rotations, and $\gamma = 0.0$ giving quartimax rotations. The implementation of these techniques is not trivial, but automatic routines are available (see, for example, the routine G03BAF in the NAG Fortran Library Routine Document).

Despite the proposed alternatives to the variance maximization assumption and the orthogonality constraint used in various RT methods, they still all share two fundamental properties with PCA: they assume that the meaningful components are linearly mixed (classical techniques are intimately linked to the linear assumtion and can not be generelized to nonlinear models) and that only second order statistics need be evaluated.

3. The Independent Component Analysis technique

In this section, we introduce the main concepts underlying the Independent Component Analysis (ICA) technique. For more details, the interested reader is referred to *Bell et al.* (1995) and *Aires et al.* (2000). The ICA technique aims to extract statistically independent components, a stronger constraint than the decorrelation requirement of the PCA.

The statistical independence of two variables, h_1 and h_2, is determined when their joint distribution can be factored:

$$P(h_1, h_2) = P(h_1) \cdot P(h_2). \quad (9)$$

This constraint involves higher-order statistics whereas the decorrelation constraint only involves second-order statistics (i.e., mean and variance). Decorrelation is equivalent to statistical independence only in the case where the quantities are Gaussian distributed, so the higher-order statistics are particularly important when the analyzed data have components with non-Gaussian distributions (*Comon* 1994). Avoiding the a priori assumption that second-order statistics are sufficient is important when the components are unknown as is usually the case.

It is also important to distinguish the non-Gaussian character of the components, $\boldsymbol{\sigma}$, with the non-Gaussian character of the data, $\boldsymbol{x}$, itself in Eq. (5). If the data have a non-Gaussian distribution, then at least one component is also non-Gaussian, since for the simplest linear mixture of Gaussian components, the distribution would be Gaussian, but a non-linear combination of Gaussian distributions could be non-Gaussian. Some previous studies examine this non-Gaussian behavior in the data (*Burgers and Stephenson* 1999; *Aires et al.* 2000).

A variable is characterized by all its statistical cumulants: the first cumulant is the mean, the second cumulant is the variance, the third cumulant is the skewness, the fourth cumulant is the kurtosis, etc (*Press et al.*, 1992). For Gaussian variables, cumulants higher than 2 are zero. When data have zero-mean, the "skewness" $\text{skew}(X) = \frac{<X^3>}{\sigma^3}$ and the "kurtosis" $\text{kurt}(X) = \frac{<X^4>}{\sigma^4} - 3$. These cumulants are often used to test a departure from Gaussian behavior. The skewness measures the symmetry of the probability distribution function: when the skewness is positive, larger events are more probable then smaller events, and the reverse is true when the skewness is negative. The kurtosis is a measure of the sharpness of the distribution: a negative kurtosis indicates that the distribution has a broaderh central peak and larger tails than a Gaussian distribution (sub-Gaussian), a positive kurtosis indicates that the distribution has a sharper central peak (super-Gaussian distribution). The non-Gaussian character of a variable is intimately linked to nonlinear dynamics (*Palmer* 1999). For example, a nonlinear dynamical system with two atractors can result in binomial distributions. Thus, without a priori information on the Gaussianity of components in an analysis of geophysical time series, the use of ICA is recommended since its requirement of statistical independence is more general than the decorrelation assumption.

The time series observations are gathered into a dataset, $\boldsymbol{X}_j{}^t$, of M observations, $\boldsymbol{x}(j) = (x_j{}^t \ ; \ t = 1, \ldots, N)$, with $j \in \{1, \ldots, M\}$, where M is the spatial dimension of the time series and N is its temporal dimension. The times series, $\boldsymbol{x}(j)$, is assumed to be a mixture of statistically independent components $\boldsymbol{\sigma} = \{\sigma_i \ ; \ i = 1, \ldots, Q\}$:

$$\boldsymbol{x}(j) = \mathcal{G}(\boldsymbol{\sigma}(j)) \quad (10)$$

where $\mathcal{G}$ is an unknown mixture operator, which is, by hypothesis, non-singular (i.e. it can be inverted).

The goal of ICA is to retrieve a function $\mathcal{J} : \boldsymbol{x} \rightarrow \boldsymbol{h}$, where $\boldsymbol{h}$ is an estimate of $\boldsymbol{\sigma}$ and the terms $\{h_i \ ; \ i = 1, \ldots, Q\}$ are statistically independent. The estimate, $\boldsymbol{h}$, is defined as a deterministic function (linear or not)

of the observations:

$$h_i = \mathcal{J}_i(\boldsymbol{W}_i, \boldsymbol{x}) \; ; \quad i = 1, \ldots, Q \qquad (11)$$

where $\{\boldsymbol{W}_i \; ; \quad i = 1, \ldots, Q\}$ is the set of parameters of $\mathcal{J}$. As in RT, the number of components, Q, is here supposed to be known. This number can be estimated, in easy cases, by a break in the frequency spectrum of the data; for more difficult spectra, see for example (*Joliffe* 1986). With real observations, Q depends on the analysis objectives: extracting a lot of components allows for more complete description of the variability but makes the interpretation much more complicated, whereas extracting fewer components focuses attention on fewer different phenomena at the cost of explaining less of the variability. The reader interested in this topic should refer to an article by *Nadal et al.* (2000).

The parameters, $\boldsymbol{W}_i$, are estimated by applying a gradient descent algorithm to a cost function that specifies the statistical independence of the $\{h_i \; ; \quad i = 1, \ldots, Q\}$. Different equivalent cost functions can be used; we use here the *infomax* approach to ICA (*Nadal and Parga* 1994) from which simple algorithms have been derived (*Bell and Sejnowski* 1995). Information theory is used to specify the statistical independence cost function: the fundamental quantity is *information redundancy*. Given Q variables, $h_1, h_2, \ldots, h_Q$, the information redundancy, $\mathcal{R}(h_1, h_2, \ldots, h_Q)$, is defined as the Kullback divergence (*Dacunha-Castelle and Duflo* 1982) between the joint distribution, $P_h(h_1, h_2, \ldots, h_Q)$, and the factorized distribution, $P_1(h_1) \cdot P_2(h_2) \ldots P_Q(h_Q)$:

$$\mathcal{R}(\boldsymbol{h}) = \int_{-\infty}^{+\infty} \prod_{i=1}^{Q} dh_i \; P_h(\boldsymbol{h}) \; \log \frac{P_h(\boldsymbol{h})}{\prod_{i=1}^{Q} P_i(h_i)} \qquad (12)$$

This information redundancy measures the difference between the joint and the factorized distribution: when the redundancy $\mathcal{R}(\boldsymbol{h}) = 0$, $P_h(\boldsymbol{h}) = \prod_{i=1}^{Q} P_i(h_i)$, which means, by the definition in equation (9), that the components of vector $\boldsymbol{h}$ are statistically independent. An important remark is that, since no geometric constraint on the basis functions is specified in the redundancy quality criterion of (12), the base functions extracted by ICA, in contrast to PCA, can be non-orthogonal.

A statistical regression model for the extraction model in Equation (11) has to be specified. For the nonlinear mixture case the regression model needs to be nonlinear in order to simulate $\mathcal{G}^{-1}$. The Multi-Layer Perceptron (MLP), an artificial Neural Network model, could be chosen for such a case. The nonlinear mixture case will be the subject of future work.

In that present linear mixture case, we use a simpler MLP architecture with no hidden layers as the extraction model. This neural mapping is defined by (from right to left in Figure 2):

$$\boldsymbol{y} = f(\boldsymbol{h}) = f(\boldsymbol{J} \cdot \boldsymbol{x}), \qquad (13)$$

where f is the logistic function (nonlinear, bounded, and invertible). This model is basically a linear model, but it employes a nonlinear logistic function. A nonlinear f is used, even if the mixture model is linear, for algorithmic considerations (*Nadal and Parga* 1994): to obtain statistical independence, the manipulation of higher moments (like $< {h_i}^3 >, < {h_i}^4 >, < {h_i}^5 >, \ldots$) is required. Applying nonlinear f_i to the h_i's allows one to include these higher-order moments because the Taylor expansion of $f(\boldsymbol{h})$ uses higher powers of the h_i values. So we consider a nonlinear transformation (postfiltering step) on the estimator $\boldsymbol{h} = \boldsymbol{J} \cdot \boldsymbol{x}$: the extracted components are not the output $\boldsymbol{y}$ or the neural mapping (13), but the vector $\boldsymbol{h} = \boldsymbol{J} \cdot \boldsymbol{x}$.

The parameters, $\boldsymbol{W}_i$, defining the matrix $\boldsymbol{J}$ and the optimal transfer functions f have to be determined by minimizing the redundancy criterion in (12). Practically, it has been demonstrated in various applications (*Bell and Sejnowski* 1995) that full optimization of the transfer functions is not necessary for performing ICA. Although promising results have been obtained, this analysis strategy can be improved by introducing some partial adaptation of the transfer functions to that particular problem. We use here the classical sigmoide function $f(x) = 1/(1 + e^{-\beta * x})$ that has proven generally usefull.

With the information redundancy reduction criterion, (12), and a no-hidden-layer architecture, a straightforward algorithmic implementation of the ICA has been found (*Bell and Sejnowski* 1995) to estimate the matrix $\boldsymbol{J}$:

$$J_{ik}(n+1) = J_{ik}(n) + \rho(n)(J_{ik} + \tilde{y}_i \cdot \sum_l J_{lk} \cdot h_l) \qquad (14)$$

where $J_{ik}(n)$ is an element of the matrix $\boldsymbol{J}$ at step n of the gradient descent, ρ is the learning rate parameter of the gradient descent, and

$$\tilde{y}_i = \frac{\partial}{\partial y_i} \frac{\partial y_i}{\partial h_i} = \frac{\partial}{\partial h_i} ln(\frac{\partial y_i}{\partial h_i}) = 1 - 2y_i. \qquad (15)$$

This algorithm is described in a more practical way

in the appendix [2]. Note that, although the theory behind this analysis method may seen complex, the actual computational procedure that results for the linear case is relatively simple.

ICA can be applied to the raw data, $\boldsymbol{x}(j)$, but it has been shown (*Nadal et al.* 2000; *Aires et al.* 2000) that a PCA pre-processing of observations makes the gradient descent step stabler and faster. The N-dimensional data, $\boldsymbol{x}(j)$, is projected onto its first N' (where $N' < N$) principal components using the matrix $\boldsymbol{J}_0$ of PCA: observation noise is reduced and fewer free parameters need to be estimated because the dimensions of the matrix $\boldsymbol{J}$ are considerably reduced, from $Q \times N$ to $Q \times N'$ The dimension of the compression N' is chosen to be equal to Q, the number of extracted components, thus, $\boldsymbol{J}$ is a $Q \times Q$ matrix. The PCA-compression pre-processing step does not alter the components extracted by ICA if the negligected components (including noise) are the statistically weak sources (see *Nadal et al.* 2000 for a definition of the term "weak"). In this configuration, the ICA can be considered equivalent to performing an oblique rotation of the PCA solution by generalizing decorrelation to statistical independence, but no additional criterion must be selected from among a large number of possibilities like the localization constraint in classical RT. The rotation matrix of the PCA solution $\boldsymbol{J}_0$ is the ICA solution $Q \times Q$ matrix $\boldsymbol{J}$. If the real physical components are Gaussian-distributed, then decorrelation is sufficient and the ICA algorithm would not change the PCA solution since zero-gradient is already reached. In the case where the components are non-Gaussian, the ICA rotates the initial PCA solution. So ICA improves the PCA solution only in the non-Gaussian case. We note that there are many examples in the literature showing non-Gaussian distributions of climate parameters (e.g., cloud optical thickness (*Rossow and Schiffer* 1991), precipitation water path (*Lin and Rossow* 1996) abd SST (*Aires et al.* 2000)).

[2] See also the web page http://www.cnl.salk.edu/~tewon/ica_cnl.html of the Computational Neuroscience Laborabory of Terry Sejnowski at The Salk Institute for links to recent literature, software and demos concerning the ICA paradigm.

4. Application to a linear sum of components

a. Construction of the synthetic dataset

Geophysical time series have been analyzed by linear statistical extraction techniques for decades. The synthetic dataset used in this study is generated to mimic the apparent expectations of such an analysis approach, namely, that the observations are a linear sum of modes with very different space and time variations and, so, are separable by such an analysis. However, we also include two modes that are spatially overlapped, but with different time behaviors, and two spatially separated modes with the same time behavior representing a teleconnection. We select $Q = 6$ components representing six different dynamical phenomena, each described by a different temporal basis function, $\boldsymbol{g}_i$ (solid lines in Figure 3), constructed from composites of sinusoids with different frequencies and phases. Each basis function has been normalized to give a temporal standard deviation of unity. The temporal dimension of these basis functions is taken to be $N = 365$ (e.g., one year of daily data). A spatial resolution of $2.5^0 \times 2.5^0$ is chosen, corresponding to $M = 144 \times 72 = 10368$ pixels. Finally, the dataset, $\boldsymbol{X}_j{}^t = \{x(j) \in R^N \; ; \; j = 1, \ldots, M\}$, where R^N is the space of real vectors of dimension N, $N = 365$ and $M = 10368$, is formed from the time series $x(j)$ for each pixel j by the linear sum of the basis functions, $\boldsymbol{x}(j) = \boldsymbol{g}_1\sigma_1(j) + \ldots + \boldsymbol{g}_Q\sigma_Q(j) + \varepsilon$ (linear model of Eq. 5). The term ε is Gaussian-distributed noise (zero mean and standard-deviation of 0.5), representing very noisy data as might be the case when analysing climate anomalies.

The $\{\sigma_i(j) \; ; \; i = 1, \ldots, Q\}$ indicate the strength of each component, i, at each pixel, j, i.e., the spatial distributions. These strengths are constructed to have a geographical Gaussian distribution, giving a different ellipsoidal distribution for each component (left column in Figure 4). Artificial land contours are introduced in the display of σ_i for easier description of the modes. One of the components has two peaks in its spatial distribution (near the Americas) to represent a teleconnection pattern (map of component 1 in left column of Figure 4), so the total number of ellipsoidal peaks is seven. Also, the geographical extent of two of the components overlaps in the Indian Ocean (maps of components 4 and 6 in left column of Figure 4) to complicate the component extraction process.

The variance contributed by the $Q = 6$ compo-

nents and the added noise are shown in Table 1: the components produce 67 % of the total variance and the noise produces 33 %. The total variance of a component results from the combination of the temporal variability of the basis function (as a function of normalized amplitude and frequency) and the spatial extent of the component.

b. Results of PCA and ICA

The PCA components are determined by computing the matrix $\boldsymbol{J}_0$. The best number of PCA components to extract is determined here by observing the spectrum of cumulative percent of variance explained by the PCA components (Figure 5). More sophisticated criteria have been developed to determine the number of significant components (see for example (*Joliffe* 1986)). The first $Q = 6$ PCA components represent 67.7 % of the total variance and the 359 remaining components explain equal portions of the remaining 32.3 % of the total variance, representing the noise in the dataset. The PCA temporal basis functions (crossed lines in Figure 3) are each compared with the real basis function to which it best corresponds. PCA basis functions 2, 3 and 4 provide a relatively good estimate of the true functions, although there are some errors near the peak values. PCA temporal basis functions 1, 5 and 6 (low frequencies) are much worse fits. In particular, higher frequencies have been mixed into the real basis functions.

The corresponding PCA component maps are defined at pixel j by the values $(\sigma_1, \ldots \sigma_Q)(j) = \boldsymbol{J}_0 \cdot \boldsymbol{X}_j{}^\star$ and are shown in Figure 4 (middle column), where $\boldsymbol{X}_j{}^\star$ is the j^{th} column of data matrix $\boldsymbol{X}$. We see that the PCA (or EOF) technique confuses elements from the different components, the general mixing problem, such that all of its components exhibit many more geographic peaks than in the real components. Even if the corresponding PCA temporal basis function is relatively well retrieved, the corresponding component map still exhibits the mixing problem (see expecially PCA basis function 2 in Figure 3 and the corresponding PCA component map in Figure 4). One cause of the mixing is well illustrated in Table 1 where the variance explained by each PCA component is compared to the variance of the actual components. The first PCA component explains 24.4 % of the total variance, which is much more than its true variance of 13.3 %. The 6^{th} PCA component represents only 3.1 % which is a considerable underestimate of the real value of 10 %. Thus, the variance maximization constraint on the solution in PCA shifts signal from other components into the first component, producing a mixture of many true component variabilities. The noise level estimate of 32.3 % is a good estimate, but its small under-estimate of the real noise is due to the projection of noise into the first 6 PCA components (representing 0.7 %).

Particularly notable in Figure 4 is that if not rotated the mixing tendency of the PCA could suggest many more teleconnections in observations than are actually present. Since in this synthetic case all six components contribute roughly the same amount of variance (10-13 %), the PCA technique has combined many of the actually-separate components into several of its components, trying to maximize the amount of variance explained by each. However, the method is then compelled to alternate positive and negative values to compensate for having too much variance when the components are added back together. This effect is especially apparent for the overlapping components in the Indian Ocean: two PCA basis functions possess broad central peaks spanning the geographic distribution of both of the real components and two others possess, in this same location, two opposite-signed peaks (see PCA component maps of components 1,4, 5 and 6 in Figure 4, middle column). The alternating positives and negatives in PCA are partly the result of the orthogonality constraint. A similar projection of real components into more than one PCA component occurs when a geographically isolated mode moves during the time period (*Kim and Wu* 1999). Moreover, the component with two peaks near the Americas, representing a real teleconnection, shows up in four of the PCA components (components 1, 3, 4 and 6 in Figure 4, middle column), but mixed with other components as well, suggesting teleconnections between the Americas and the South Atlantic and Indian Oceans that do not exist.

ICA can be applied directly to the raw data, $\boldsymbol{x}(j)$, but, as previously commented in *Nadal et al.* (2000) and more briefly at the end of section (), a PCA pre-processing of observations makes the gradient descent step numerically stabler and faster. So the observed data, $\boldsymbol{x}(j)$, are first projected onto the first $Q = 6$ PCA components using the matrix $\boldsymbol{J}_0$. The ICA technique is then applied to the pre-processed data, $\tilde{\boldsymbol{x}}(j) = \boldsymbol{J}_0 \cdot \boldsymbol{x}(j)$ (dimension $Q = 6$ instead of $N = 365$). As explained in section (), this is equivalent to performing a rotation on the PCA initial solution where the rotation matrix is the $Q \times Q$-matrix $\boldsymbol{J}$ of the ICA solution. Thus the 6 ICA extracted com-

ponents explain the same amount of variance as the 6 PCA components (67.7 %).

A varimax rotation of the PCA solution (a RT method) has also been obtained for this case. Results (not shown) are improved for some of the components since the real components are geographically well localized, but the mixing problem still remains.

The six ICA basis functions are shown in Figure 3 (dotted lines). The ICA basis functions are very similar to the real basis functions. This comparison shows how the ICA technique has corrected its first guess (the PCA solution) to be closer to the true solution. The additional information obtained from the requirement for statistical independence is nicely illustrated: the ICA technique has transformed the PCA initial solution for a better retrieval of all six components. The ICA component maps are presented in Figure 4 (right column). Generally, the components are well-retrieved and separated even the teleconnection mode (ICA component 1 in Figure 4, right column) and the two overlapping modes in the Indian Ocean (ICA component 4 and 6 in Figure 4, right column). The transformation of the PCA component maps by ICA is always an improvement.

An experiment was conducted with the same data but without the noise (not shown). The ICA separates the original six modes almost perfectly and the ICA solution is very close to the real solution. This result indicates that the presence of measurement noise in a dataset will produce a small amount of mode mixing even in the ICA solution; however the results shown here (small hints of other modes in ICA component 4 and 6 in Figure 4, right column) is produced by a situation where the signal to noise ratio is only about two. Although this situation may be relevant to climate studies, ICA can separate most of the noise into its own statistically independent mode.

Table 1 shows that the variance explained by the ICA components is much closer to the real solution than the initial PCA components: the variance explained by the first couple of modes decrease and that retained by the remaining modes increase. Differences between the true and ICA explained variance for each component are less than 0.6 %, where the discrepancies are the result of the projection of some part of the noise into the ICA components.

5. Concluding remarks

For extraction of physically meaningful modes from observations, where the characteristics of the system's dynamics are not (well-) known, identifying statistically independent variation modes seems to be a sensible alternative for the rotation of a first PCA (i.e. EOF) solution which is sensitive to the mixing problem. Our simple example shows that in the most general, though still linear case (the most favorable condition for PCA), PCA will mix modes of comparable magnitude, generating spurious regional overlaps or teleconnections where none exist or distorting existing overlaps or teleconnections. We have shown the potential of the ICA technique for separating a complex signal in a more meaningful way. The mixing problem inherent in the PCA technique and the artifacts produced by the orthogonality and maximum-variance constraints of PCA are avoided when rotated by ICA. Moreover, the use of higher-order statistics by ICA to determine statistical independence assumes only the generalization of the decorrelation used in all classical approaches. Nevertheless, even statistical independence does not guarantee that the modes produced by different physical processes will be separated.

ICA, by finding statistically independent modes, may provide a better way to explore the unknown dynamics of a system. In the case of climate variations, where the components of the system are probably coupled (see for example, *Salby and Callaghan* 2000 or *Krishnamurthy and Goswami* 2000), considering the modes to be as statistically distinct as possible, even with a linear-ICA, would provide "prototypical" components that might serve as a guide to further investigation. As with the classical PCA technique or classical RT, this first (linear) ICA algorithm is not able to deal correctly with propagating components or components mixed nonlinearly. However, the ICA paradigm (statistical independence) may be a sufficiently powerful concept to be generalized using more advanced statistical models (e.g., more complicated neural networks) to treat nonlinear problems. This requires development of nonlinear solution algorithms and their testing for cases where the combination of modes is nonlinear, when components are physically linked, and for cases with propagating modes.

Appendix: Principal steps of the algorithm

We adopt here the linear model $\boldsymbol{x} = \boldsymbol{G} \cdot \boldsymbol{\sigma}$, where $\boldsymbol{x}$ is the observation, $\boldsymbol{G}$ is the basis function matrix and $\boldsymbol{\sigma}$ is the vector of components to estimate. The goal of the statistical decomposition technique

is to estimate a matrix $\boldsymbol{J} = \boldsymbol{G}^{-1}$ (the superscript $^{-1}$ represents the pseudo-inverse if $\boldsymbol{G}$ is not square), the filter matrix, using only a dataset of observations $\{\boldsymbol{x}^e \ ; \ e = 1, \ldots, E\}$, where E is the number of samples in the dataset. With thr matrix $\boldsymbol{J}$ applied to each observation $\boldsymbol{x}$, the components $\boldsymbol{\sigma}$ are estimated by $\boldsymbol{\sigma} \simeq \boldsymbol{h} = \boldsymbol{J} \cdot \boldsymbol{x}$, and the basis function matrix $\boldsymbol{G}$ is estimated by the inverse matrix $\boldsymbol{J}^{-1}$.

The principal steps of the time series analysis by the ICA technique are:

• **Optional pre-processing**: The dataset $\boldsymbol{X_j}^t = \{\boldsymbol{x}(j) \in R^N \ ; \ j = 1, \ldots, M\}$, where t is the time index and j is the space index (geographical locations), may require pre-processing: (1) spatial, temporal or spatio-temporal interpolation to fill in missing data, (2) filtering of data to suppress undesirable frequencies (noise effects), (3) de-trending to obtain stationary data, and (4) removing the annual cycle to examine interannual anomalies. None of these steps is required.

• **Chose the space for the decomposition**: (1) in time, which is the approach we have adopted in our study:

$$\boldsymbol{x}(j) = \boldsymbol{g}_1 \sigma_1(j) + \ldots + \boldsymbol{g}_Q \sigma_Q(j) + \boldsymbol{\varepsilon},$$

(2) in space, (3) in frequency, or (4) in a mixture of these spaces. The observations (a time series or a geographical field, ...) are denoted, in the following, by the d-dimensional vector $\boldsymbol{x}^e$ and the dataset is $\{\boldsymbol{x}^e \ ; \ e = 1, \ldots, E\}$.

• **Center the dataset**: The observation mean $< \boldsymbol{x}^e >$ is removed from the dataset: $\boldsymbol{x}^e \leftarrow \boldsymbol{x}^e - < \boldsymbol{x}^e >$. This step is necessary for statistical techniques where data are supposed to have zero-mean like ICA.

• **Optional normalization**: If the user wants to put the same statistical weight on each coordinate of the observation $\boldsymbol{x}^e$: then the dataset can be normalized by the standard-deviation vector $\boldsymbol{x}^e \leftarrow \boldsymbol{x}^e / \boldsymbol{e}_x$.

• **Optional Eigen-vector decomposition**: The covariance (or correlation, in the case of normalized observations) matrix $< \boldsymbol{x}^t \cdot \boldsymbol{x} >$ is estimated from the dataset. The eigenvalues $\boldsymbol{\Lambda}$ (diagonal matrix) and the eigen-vector matrix $\boldsymbol{V}$ of $< \boldsymbol{x}^t \cdot \boldsymbol{x} >$ are then computed using a classical numerical routine. The number of PCA or ICA extracted components Q is chosen by observing the spectrum of eigenvalues.

• **Optional PCA solution**: The PCA solution is computed to pre-process the data:

- The $d \times Q$ PCA basis function matrix, $\boldsymbol{G_{PCA}}$, contains in its columns the first Q eigen-vectors of $\boldsymbol{V}$ (the columns of $\boldsymbol{V}$ represent time series in the time decomposition, and geographical field in the space decomposition, ...).
- Since, by definition, $\boldsymbol{V}^{-1} = \boldsymbol{V}^t$, the filter PCA matrix, $\boldsymbol{J_{PCA}}$, is equal to the transposed $Q \times d$ basis function matrix, $\boldsymbol{G_{PCA}}$. Then, the extracted components, $\boldsymbol{h}$, that estimate the true components, $\boldsymbol{\sigma}$, are the projection of the observations, $\boldsymbol{x}$, onto the filters: $\boldsymbol{h} = \boldsymbol{J_{PCA}} \cdot \boldsymbol{x}$.
- The first Q eigenvalues in $\boldsymbol{\Lambda}$ represent the variability explained by each of the Q components.

• **ICA solution**:

1 Pre-whitening of dataset: The PCA solution is used as a pre-processing step: the observations $\boldsymbol{x}^e$ are projected onto the PCA filters: $\boldsymbol{x}^e \leftarrow \boldsymbol{J_{PCA}} \cdot \boldsymbol{x}^e$. The ICA algorithm is then applied into these Q-dimensional data.

2 The ICA solution, $\boldsymbol{J_{ICA}}$, is initialized as the identity matrix $\boldsymbol{I_{Q \times Q}}$. This, associated to the previous whitening step, is equivalent to taking the PCA solution as first guess for ICA.

3 For the minimization of the criterion specifying the statistical independence, a gradient descent algorithm is used. The classical gradient descent uses all the samples of the dataset to compute a mean ΔJ_{ik} in Equation (14). This algorithm is called the deterministic gradient descent. The major inconvenience of this algorithm is that it can be trapped in local minima. We use, in our applicaton, the stochastic gradient descent algorithm that uses the gradient descent formula (14) iteratively in unique random samples of the dataset. The stochastic character of the optimization algorithm allows theoretically, and under some constraint not discussed here, for the optimization technique to reach the global minimum of the criterion instead of a local minimum (*Duflo* 1996).

4 An observation $\boldsymbol{x}^e$ is randomly chosen in the dataset. The propagation through the neural network (chosen model for the component extraction) is given by: $\boldsymbol{y} = f(\boldsymbol{h}) = f(\boldsymbol{J_{ICA}} \cdot \boldsymbol{x}^e)$, where $f(a) = [1 + exp(-\beta \cdot a)]^{-1}$ is the logistic function (β is a parameter controlling the slope of the logistic function, we take $\beta = 2.0$ as in (*Bell and Sejnowski*, 1995). The FORTRAN

routine of this process is:

c - - propagation through the neural network

do $i = 1, d$

$h(i) = 0.d0$

do $k = 1, d$

$h(i) = h(i) + J_{ICA}(i,k) * x^e(k)$

enddo

$h(i) = h(i) + bia(i)$

$y(i) = 1.d0/(1.d0 + dexp(-\beta * h(i)))$

enddo

where *bia* is the classical bias vector in a MLP neural network (not shown in the text for simplicity). We use, in this routine, double precision variables to avoid numerical instabilities.

5 The learning process is then defined as:

c - - transitory quantities

do $j = 1, d$

$hhh(j) = 0.d0$

do $k = 1, d$

$hhh(j) = hhh(j) + J_{ICA}(k,j) * h(k)$

enddo

enddo

c - - modification of weights

do $i = 1, d$

do $j = 1, d$

$J_{ICA}(i,j) = J_{ICA}(i,j) + param*$

& $(J_{ICA}(i,j) + \beta * (1.d0 - 2.d0 * y(i)) *$ $hhh(j))$

enddo

$bia(i) = bia(i) + \beta * (1.d0 - 2.d0 * y(i))$

enddo

where *param* is the learning parameter of the gradient descent optimization (we take $param = 0.0005$).

6 Stopping criterion: many criteria can be used to define when to stop the learning cycle. The simplest criterion is to determine a priori the number of learning steps. A better criterion is to determine when the difference between solution $\boldsymbol{J}_{ICA}$ at time t and at time $t+1$ falls below some threshold value. Another stopping criterion is to evaluate the statistical independence of the extracted components $\boldsymbol{h}$: cumulants (i.e. additive higher-order moments) are a practical way to do that, but this approach is computationally expensive. The learning algorithm returns to step 4 until the stopping criterion is reached.

• **Analysis of results**: When the matrix $\boldsymbol{J}_{ICA}$ has been determined by ICA, the global ICA filters (taking into account the PCA pre-processing) are defined by the $Q \times d$ matrix: $\boldsymbol{J}_{GLO} = \boldsymbol{J}_{ICA} \cdot \boldsymbol{J}_{PCA}$

- The projection of data is used to estimate the components: $\boldsymbol{h} = \boldsymbol{J}_{GLO} \cdot \boldsymbol{x}^e$
- The $d \times Q$ ICA basis function matrix $\boldsymbol{G}_{GLO} = \boldsymbol{J}_{GLO}{}^{-1} = \boldsymbol{G}_{PCA} \cdot \boldsymbol{J}_{ICA}{}^{-1}$ is normalized to obtain normalized ICA basis functions, as in PCA approach.
- Computation of explained variance of each of the basis functions.

Acknowledgments. We are grateful to Dr. David Rind and Dr. Ronald L. Miller for their helpful comments. This work was supported by special funding provided by Dr. Robert J. Curran, NASA Climate and Radiation Program.

References

Aires F., A. Chédin, and J.-P. Nadal, 2000: Independent Component Analysis of Multivariate Times Series: Application to the tropical SST variability, *Journal of Geophysical Research*, **105**, D13, 17,437–17,455.

Atick, J. J., Could information theory provide an ecological theory of sensory processing, 1992: *Network Comput. Neural Syst.*, **3**, 213–251.

Bell A. J., and T. J. Sejnowski, 1995: An information-maximization approach to blind separation and blind deconvolution, *Neural Computation*, **7**, 6, 1004–1034.

Broomhead, D., and G. King, 1986: Extracting qualitative dynamics from experimental data, *Physica D*, **20**, 217–236.

Burel, G., 1992: Blind source separation of sources: A nonlinear algorithm, *Neural Networks*, **5**, 937–947.

Burgers, G., and D.B. Stephenson, 1999: The "Normality" of El Ninõ, *Geophysical Research Letters*, **26**, 8, 1027–1030.

Cardoso, J.F., and Jutten, C., and P. Loubaton (Eds), 1999: *ICA'99, First International Workshop on Independent Component Analysis and Signal Separation , Aussois, France.*

Comon, P., 1994: Independent Component Analysis, a new concept ?, 1994: *Signal Process.*, **36**, 287-314.

Dacunha-Castelle, D., and Duflo, M., 1982: Probabilités et Statistiques - Tome 1: probleèmes à temps fixe, Masson, Paris.

Duflo, M., 1996: Algorithmes stochastiques, Mathématiques et Applications, Springer.

Horel, J., 1981: A rotated principal component analysis of the interannual variability of the northern hemisphere 500 mb height field, *Monthly Weather Review*, **109**, 2080–2092.

Hyvärinen, A., and Oja, E., 2000: Independent component analysis: algorithms and applications, *Neural networks*, **13**, 411–430.

Joliffe, I.T., 1986: Principal component analysis, Springer series in statistics, 271 p., New York, Springer-Verlag.

Jutten, C., and J. Herault, 1991: Blind separation of sources, part I, An adaptive algorithm based on neuromimetic architecture, *Signal Process.*, **24**, 1–10.

Karl, T. R., Koscielny, A. J., and Diaz, H. F., 1982: Potential Errors in the Application of Principal Component (Eigenvector) Analysis to Geophysical Data, *Journal of Applied Meteorology*, **21**, 1183–1186.

Kim, K.-Y., and Q. Wu, 1999: A comparison study of EOF techniques: analysis of nonstationary data with periodic statistics *Journal of Climate*, **12**, 185–199.

Korres, G., Pinardi, N., and A. Lascaratos, 2000: The Ocean Response to Low-Frequency Interannual Atmospheric Variability in the Mediterranean Sea. Part II: Empirical Orthogonal Functions Analysis, *Journal of Climate*, **13**, 4, 732–745.

Krishnamurthy, V., and B.N. Goswami, 2000: Indian Monsoon-ENSO Relationship on Interdecadal Timescale, *Journal of Climate*, **13**, 3, 579–595

Lin, B, and Rossow, W.B., 1996: Seasonal variation of liquid and ice water path in nonprecipitating clouds over oceans, *Journal of Climate*, **9**, 2890–2902.

Lorenz, E., 1951: Seasonal and irregular variations of the northern hemisphere sea-level pressure profile, *Journal of Meteorology*, **8**, 52–59.

Monahan, A.H., 2000: Nonlinear Principal Component Analysis by Neural Networks: Theory and Application to the Lorenz System, *Journal of Climate*, **13**, 4, 821–835.

Nadal J.-P., and N. Parga, 1994: Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer, *Computation in Neural Systems*, **5**, 565–581.

Nadal J.-P., and Parga N., 1997: Redundancy reduction and independent component analysis: Conditions on cumulants and adaptive approaches, *Neural Computation*, **9**, 7, 1421–1456.

Nadal J.-P., E. Korutcheva, and F. Aires, 2000: Blind source separation in the presence of weak sources, *Neural Networks*, **13**, 589–596.

Palmer, T.N., 1999: A nonlinear dynamical perspective on climate prediction, *Journal of Climate*, **12**, 575–591.

Richman, M., 1981: Obliquely rotated principal components: an improved meteorological map typing technique ?, *Journal of Applied Meteorology*, **20**, 1145–1159.

Richman, M., 1986: Rotation of principal components, *Journal of Climatology*, **6**, 293-335.

Rossow, W.B., A.W. Walker, and L.C. Garder, 1993: Comparison of ISCCP and other cloud amounts, *Journal of Climate*, **6**, 2394–2418.

Rossow, W.B., and R.A. Schiffer, 1991: ISCCP cloud data products, *Bull. Amer. Meteor. Soc.*, **72**, 2–20.

Salby, M., and P. Callaghan, 2000: Connection between the Solar Cycle and the QBO: the Missing Link *Journal of Climate*, **13**, 2, 328–338.

Vautard, R., P. Yiou, and M. Ghil, Singular-spectrum analysis: A toolkit for short, noisy chaotic signals, 1992: *Physica D*, **58**, 95–126.

Vautard, R., C. Pires, and G. Plaut, Long-range atmospheric predictability using space-time principal components, 1996: *Mon. Weather Rev.*, **124** (2), 288–307.

Von Storch, H., and C. Frankignoul, 1998: Empirical modal decomposition in coastal oceanography, Chap. 16, in *The Sea*, vol. 10, edited by Kenneth H. Brink and Allan R. Robinson, John Wiley, New York, 419–455.

Xue, Y., Leetmaa, A., and M. Ji, 2000: ENSO Prediction with Markov Models: The Impact of Sea Level, *Journal of Climate*, **13**, 4, 849-871.

Filipe Aires and Willian B. Rossow, NASA Goddard Institute for Space Studies, 2880 Broadway, New-York, NY 10025, USA. (e-mail: faires@giss.nasa.gov; wrossow@giss.nasa.gov)

Alain Chédin, CNRS Laboratoire de Météorologie Dynamique,
École Polytechnique, 91128 Palaiseau Cedex, France. (e-mail: chedin@jungle.polytechnique.fr)

...., 2000; revised ...., 2000; accepted ...., 2000.

This preprint was prepared with AGU's LaTeX macros v4. File artifi2 formatted December 21, 2000.

Figure 1. Illustration of the Problems encountered by PCA when observations, having dimension 2 (coordinates X and Y), come from two components defining ellipses E1 and E2. The line D represents the first PCA axis defining the first PCA component: A) mixing due to the maximum-explained-variance constraint, B) indeterminacy when two components have same variance, and C) mixing due to the non-orthogonality of components.

Figure 2. The component extraction model: the perceptron architecture, where $\boldsymbol{x}$ is the observation, $\boldsymbol{h}$ is the extracted component vector and $\boldsymbol{y}$ is the network ouput.

Figure 3. Temporal basis functions, g_i: ACTUAL (solid lines), PCA estimates (crossed lines) and ICA estimates (dotted lines).

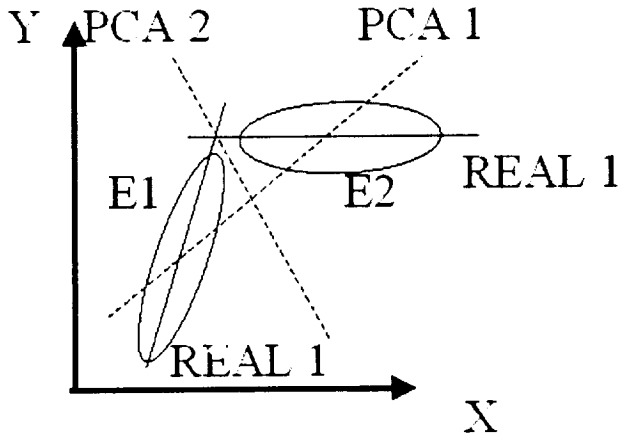
Figure 4. The maps of the actual components, σ_i (left column), of the PCA extracted components, h_i (middle column), and of the ICA extracted components, h_i (right column): components number 1-6 from the top to the bottom, component maps have been centered and normalized for comparison purposes. The continental outlines are artificial and used to make discussion of specific features easier.

Figure 5. Cumulative percent of explained variance by the PCA components.

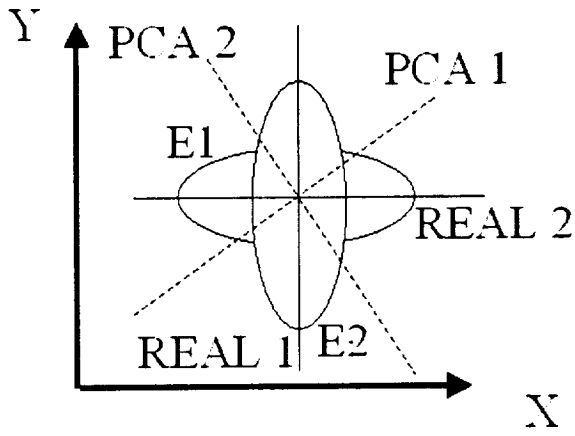
Table 1. Variance explained by Noise, REAL, PCA and ICA components

Component	REAL	PCA	ICA
1	13.3	24.4	12.7
2	12.6	14.5	13.0
3	10.7	10.7	11.3
4	10.7	8.8	10.2
5	10.7	6.2	10.3
6	10.0	3.1	10.0
Noise	33.0	32.3	32.3

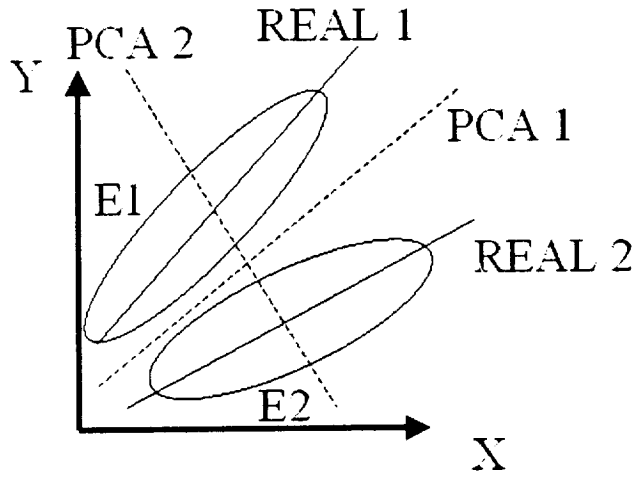
A)

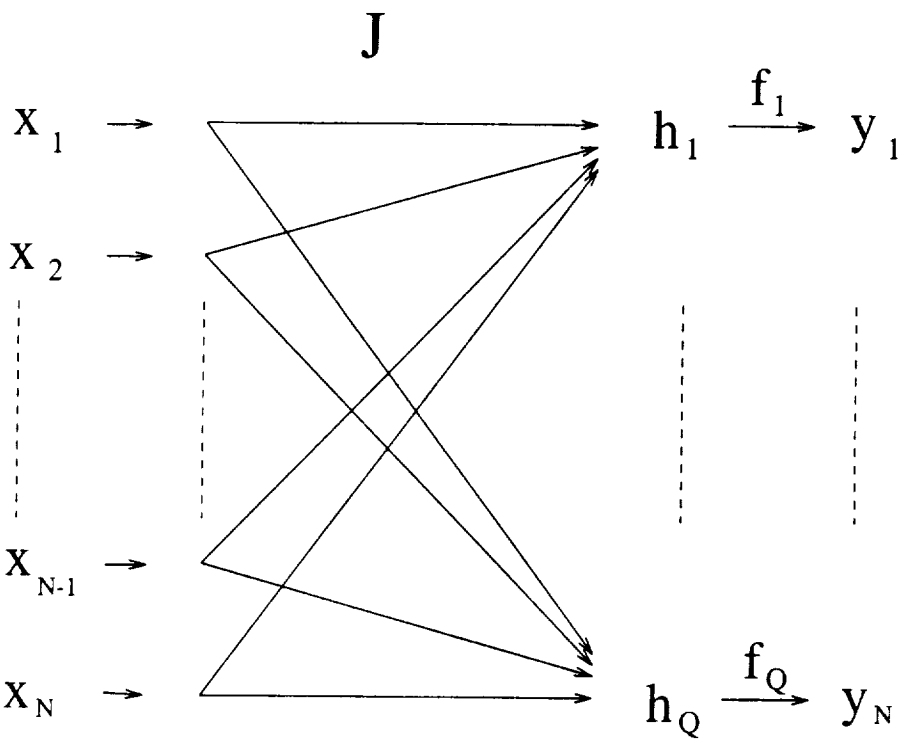

Y
PCA 2
PCA 1
REAL 1
E1
E2
REAL 1
X

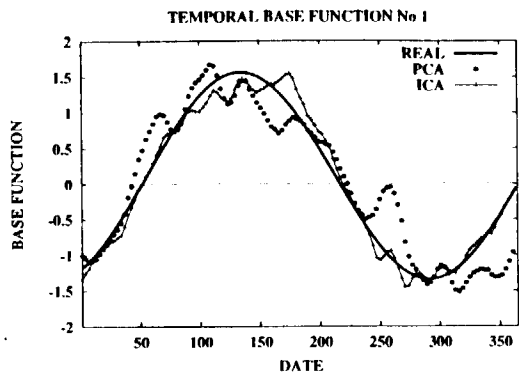
B)

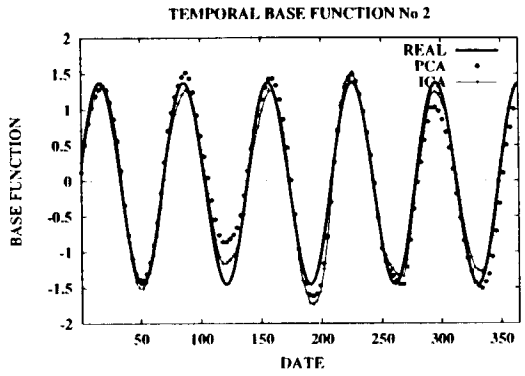

Y
PCA 2
PCA 1
E1
REAL 2
REAL 1
E2
X

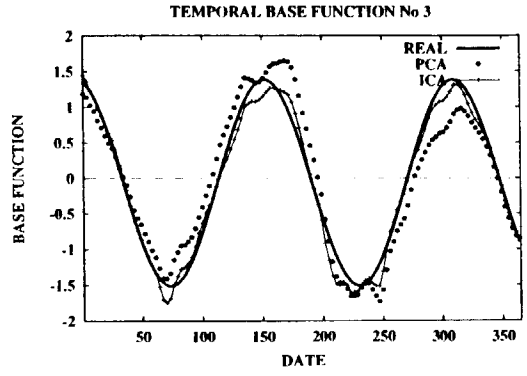
C)

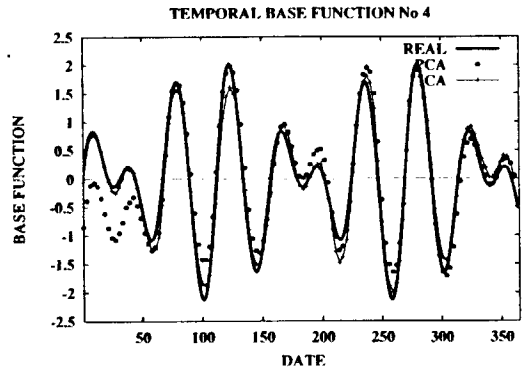

PCA 2
REAL 1
Y
PCA 1
E1
REAL 2
E2
X

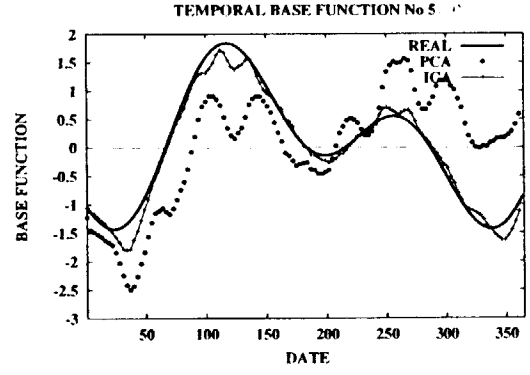

J
x 1
x 2
x N-1
x N
h 1
h Q
f 1
f Q
y 1
y N

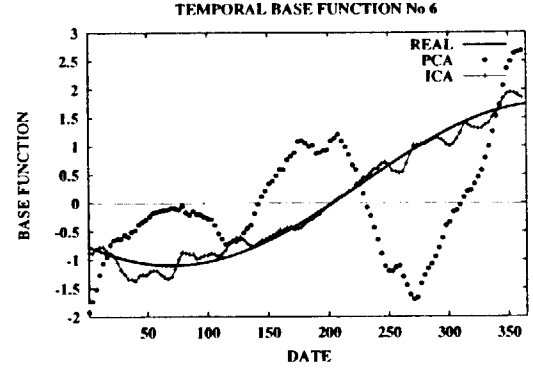

TEMPORAL BASE FUNCTION No 1
REAL
PCA
ICA
BASE FUNCTION
DATE

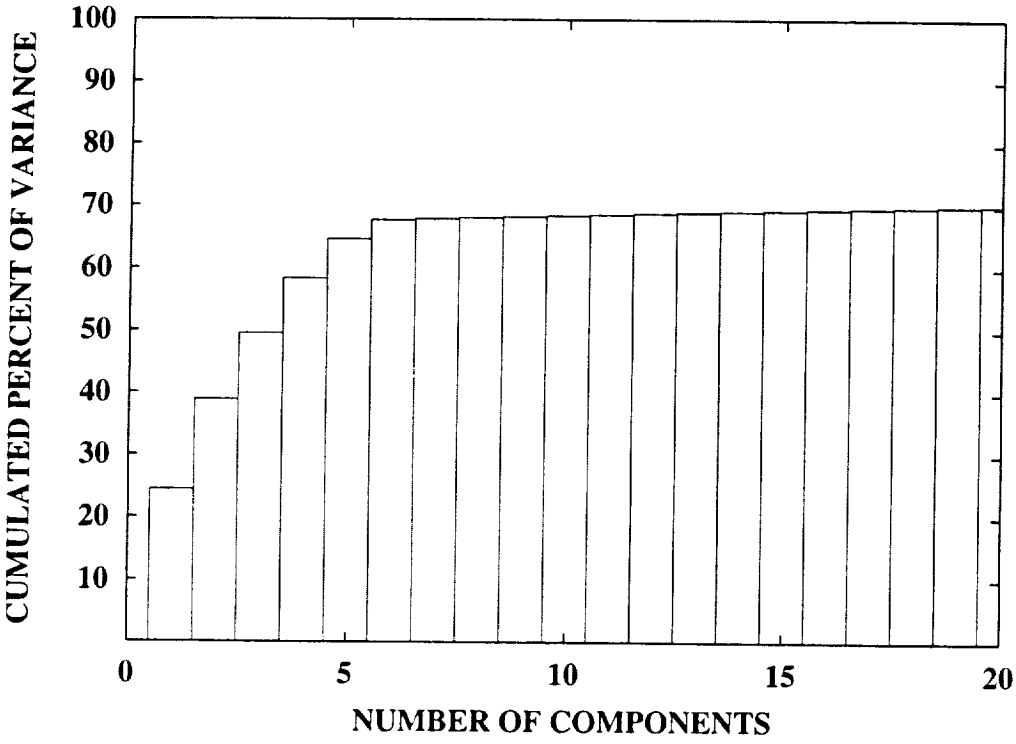

TEMPORAL BASE FUNCTION No 2
REAL
PCA
ICA
BASE FUNCTION
DATE


TEMPORAL BASE FUNCTION No 3
REAL
PCA
ICA
BASE FUNCTION
DATE


TEMPORAL BASE FUNCTION No 4
REAL
PCA
ICA
BASE FUNCTION
DATE


TEMPORAL BASE FUNCTION No 5
REAL
PCA
ICA
BASE FUNCTION
DATE


TEMPORAL BASE FUNCTION No 6
REAL
PCA
ICA
BASE FUNCTION
DATE


CUMULATED PERCENT OF VARIANCE
100
90
80
70
60
50
40
30
20
10
0
5
10
15
20
NUMBER OF COMPONENTS